RT-03f Perspectives on Metadata from Cambridge University

Sue Tranter, Marcus Tomalin & Phil Woodland

13th Nov. 2003

Cambridge University Engineering Department

Perspectives on MDE upto RT-03f

- ✓ We survived two dryruns and two evaluations!

- ✓ We have some training, dev and eval data for all tasks.

- ✓ We have generally accepted the scoring metrics and tools.

- ✗ The scoring tools are $still$ unstable, even after the evaluation submissions.

- ✗ There is still no clear 'end-user' aim.
  Are we aiming for 'readability', labelling for downstream, helping STT etc. ?

[For further details of our views see the CUED RT-03f MDE Questionnaire
response @ `http://macears.ll.mit.edu/mactech_mail/0230.html`]

Comments on RT-03 MDE evaluations

- CTS Diarisation - We would like to drop this.
  No benefit for 'readability' when single-speaker per side.
  References/scoring rules may need to be site-specific to help STT.

- BN Diarisation/SA-STT - Are we only interested in 'readability'?
  Is it important to label regions with no transcripts?
  Should diarisation include other audio events not just speakers?
  Can we score this in a way more relevant to STT?

- SU Detection - Finalising the task definition.
  We would like to expand the task to include subtypes.
  su-eval/rt-eval are not yet giving sufficiently similar numbers/trends.
  We should try to standardize and stabilize the tools quickly.

Wishlist for MDE after RT-03f

- **We need stability/continuity**
  Task definitions, Annotation Guidelines, Data and Tools should not be fundamentally changed (although minor tweaks may be desirable).

- **We need time and insight**
  Constantly moving the goal-posts does not aid research progress.
  Should we try to develop some measure of significance testing?

- **No more tasks or languages for RT-04**
  We often seem to be struggling with the number of tasks we already have.
  Large community resources should not be spent on other languages yet.

- **Management of Resources**
  If limited data/annotation time is available then English STT must get priority.
  We want more (simple) data re-using existing STT transcripts if at all possible.